Topsoil pollution forecasting using artificial neural networks on the example of the abnormally distributed heavy metal at Russian subarctic

Cite as: AIP Conference Proceedings **1836**, 020024 (2017); https://doi.org/10.1063/1.4981964
Published Online: 05 June 2017

D. A. Tarasov, A. G. Buevich, A. P. Sergeev, A. V. Shichkin, and E. M. Baglaeva

Topsoil Pollution Forecasting Using Artificial Neural Networks on the Example of the Abnormally Distributed Heavy Metal at Russian Subarctic

Tarasov D.A.[1, 2, a)], Buevich A.G.[1, b)],Sergeev A.P.[1, 2, c)], Shichkin A.V.[1, 2, d)], and Baglaeva E.M.[1, 2, e)]

[1] *Institute of Industrial Ecology UB RAS, Kovalevskoy, 20, Ekaterinburg, RUSSIA 620990.*
[2] *Institute of Radio-electronics and IT, Ural Federal University, Mira, 19, Ekaterinburg, RUSSIA 620002.*
[a)] *Corresponding author: datarasov@yandex.ru*
[b)] *bagalex3@gmail.com*
[c)] *alexanderpsergeev@gmail.com*
[d)] *and@ecko.uran.ru*
[e)] *elenbaglaeva@gmail.com*

Abstract. Forecasting the soil pollution is a considerable field of study in the light of the general concern of environmental protection issues. Due to the variation of content and spatial heterogeneity of pollutants distribution at urban areas, the conventional spatial interpolation models implemented in many GIS packages mostly cannot provide appreciate interpolation accuracy. Moreover, the problem of prediction the distribution of the element with high variability in the concentration at the study site is particularly difficult. The work presents two neural networks models forecasting a spatial content of the abnormally distributed soil pollutant (Cr) at a particular location of the subarctic Novy Urengoy, Russia. A method of generalized regression neural network (GRNN) was compared to a common multilayer perceptron (MLP) model. The proposed techniques have been built, implemented and tested using ArcGIS and MATLAB. To verify the models performances, 150 scattered input data points (pollutant concentrations) have been selected from 8.5 km2 area and then split into independent training data set (105 points) and validation data set (45 points). The training data set was generated for the interpolation using ordinary kriging while the validation data set was used to test their accuracies. The networks structures have been chosen during a computer simulation based on the minimization of the RMSE. The predictive accuracy of both models was confirmed to be significantly higher than those achieved by the geostatistical approach (kriging). It is shown that MLP could achieve better accuracy than both kriging and even GRNN for interpolating surfaces.

Keywords: Artificial Neural Networks, Chromium, Kriging, Pollution, Residual kriging

1. INTRODUCTION

All environment components (air, snow, water, soil, etc.) are known as recipients of contaminants from the multiple sources and, thus, might be utilized for studying the nature and features of the pollution (Saet et al., 1990). Rapid industrialization and human activity over the last decades has significantly contributed to the gain in soil contaminants in Arctic regions of Russia A significant heterogeneity of spatial distributions of geochemical spectra has been detected in preliminary analysis of empirical data for the various functional and geographic areas (Chukanov et al., 2006). The data being obtained in monitoring of urban territories strongly depend on relative position and intensity of emission sources as well as on building features, meteorological and hydrological conditions, climate variability and other factors. These processes and factors may cause the spatial heterogeneity and sometimes anomalies of the pollution and contaminants distributions (Zhang et al., 2008; Guo et al., 2012), such as chromium anomalies at Russian subarctic (Sergeev et al., 2010; Sergeev et al., 2015).

Applied Mathematics and Computer Science
AIP Conf. Proc. 1836, 020024-1–020024-6; doi: 10.1063/1.4981964
Published by AIP Publishing. 978-0-7354-1506-5/$30.00

Geostatistical interpolation techniques (e.g. kriging) utilize the statistical features of the measured spots together with the spatial autocorrelation between them and account for the spatial configuration of the sample spots at the prediction location. Performance of an interpolation technique might be evaluated by some statistical parameters, such as mean absolute error (MAE), root mean squared error (RMSE), and relative root mean-squared error (RRMSE).

Kriging is a common method used in spatial prediction since it estimates values for any coordinate with no bias and minimum variance (Yfantis et al., 1987; Goovaerts, 1999). Ordinary kriging weights are obtained from the kriging equation using a variogram (Matheron, 1963). The unbiased estimation of a semivariogram function (1) is a half of the RMS difference between the values of the data pairs.

$$\gamma(h) = \frac{\sum_{i=1}^{N(h)} |z(x_i) - z(x_i+h)|^2}{2N(h)} \quad (1)$$

where γ(h) is the value of a semivariogram at the distance interval h; and N(h) is the number of samples pairs at the distance interval h; $z(x_i)$ and $z(x_i+h)$ are the values for two points separated by the distance h. Considering only interpolators, which are built on the basis of weighted averages, kriging is the best interpolator with unbiased estimates, because it does not matter whether the data are normally distributed or not. It has also been frequently used for elevations (Bao et al., 2007) and soil contaminants and organic matter prediction (Dai et al., 2014; Zeissler & Hertwig, 2011). Kriging has shown considerable advantages in the prediction of soil properties, compared with deterministic methods (Schloeder et al., 2001; Liu et al., 2008; Worsham et al., 2010). The accuracy of kriging techniques depends on the density and size of sampling sites, as these methods are based on interpolation, which requires some data as inputs. At times, there is no way to get the required amount of samples at the research site. Moreover, it is found that sometimes due to the presence of significant spatial trend the stationary assumptions are violated that leads to poor interpolation results. Therefore, a more efficient method is required to improve the accuracy of interpolation methods for producing high-resolution distribution maps.

Nowadays, the famous suitable technique is artificial neural networks (ANNs). A core mathematical model of the biological neuron was established by McCulloch & Pitts (1943). ANNs provide a variety of powerful techniques for solving problems in prediction and forecasting of different entities, pattern recognition, data analysis, control and many others. They are ahead of many other methods in terms of accuracy and speed. The ability to learn makes them indispensable in solving non-standard tasks and dynamically changing challenges. In conventional MLP model, the spatial coordinates are used as the inputs and the predicted contents are used as the outputs. The functional relationship between the inputs and the outputs are established through a network of synaptic weights. These weights are determined through the learning process using iterative procedures, which sometimes takes a lot of time. To avoid the problem of local minima leading to a non-optimal solution during the learning process, some optimization algorithms applicable. The most widely used one is the Levenberg-Marquardt training method (Shepherd, 1997). The review on pattern recognition (Bishop, 1995) has detailed discussion on this subject. Artificial intelligence methodologies can help to forecast the pollutants in complicated non-linear contexts. The predictive accuracy obtained by ANNs is often higher than that of other methods or prediction of experts (Guo et al., 2012). The most frequently used ANN in environmental studies is multilayered perceptron (MLP). Due to the wide distribution, this type of network is well developed and has shown its high performance. Perceptrons are widely used among research on chemical elements distribution in soil (Dai et al, 2014; Falamaki, 2013; Li et al., 2011), in particular, heavy metals, such as Cr (Sirven et al., 2006; Anagu et al., 2009). Lots of researchers have explored MLPs for resource estimation (Samanta et al., 2004; Samanta et al., 2005; Koike et al., 2002) and proved the superiority of the MLP models over the geostatistics.

The generalized regression neural networks (GRNN) also used as interpolators and are known as universal function approximators, which can learn to approximate any continuous nonlinear function between sets of inputs and outputs (Mohanty & Majumdar, 1999; Shen et al., 2004). GRNN is a variation of the radial basis functions neural networks, which is based on kernel regression networks. GRNN does not require the learning process using iterative procedures as back propagation networks (MLP). It approximates a function drawing the estimates directly from the learning data set minimizing the estimation error by enlargement a learning data set

In this work, we propose neural networks models incorporating the techniques of generalized regression neural network (GRNN) compared to a common used multilayered perceptron (MLP) model. We examine the results obtained by applying the model to predict the pollutant levels at a particular location in the subarctic Novy Urengoy, Russia. The performance of models was evaluated by RMSE (2). The models predictive quality was assessed during the comparison between models evaluations and kriging prediction by MAE (3) and RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{iMod} - x_i)^2}{n}} \quad (2)$$

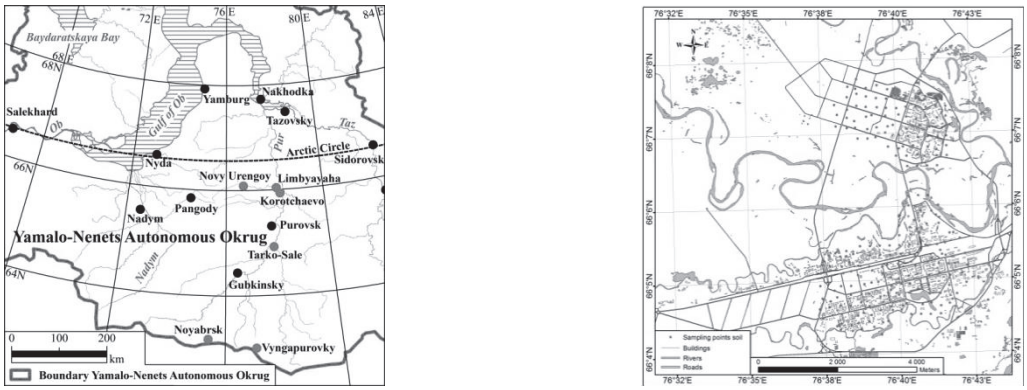
$$MAE = \frac{\sum_{i=1}^{n} |y_{Modi} - y_i|}{n} \quad (3)$$

2. MATERIALS AND METHODS

Data for the study were obtained from the results of the soil survey in Novy Urengoy, Yamalo-Nenets Autonomous Okrug, Russia (Sergeev et al., 2013; Sergeev et al., 2015), where a chromium anomaly was described. The area of sampling was approximately 8.5 km^2 (see Figure 1). In total, 150 samples were collected. Concentration indicators for the element (Cr) were obtained by chemical analysis. The descriptive statistics of the modeled element are shown in Table 1.

TABLE 1. Descriptive statistics of the modeled element (*Cr*)

Pollutant	Min	Max	Mean	SD	CV	Skewness	Kurtosis	Median
Cr	25.8	1265.4	245.2	256.3	382.9	1.41	4.61	89.5



(a) (b)

FIGURE 1. The sampling place: a) Yamalo-Nenets Autonomous Okrug, Russia; b) Novy Urengoi city

The entire data set was divided into two groups: 70% (105 samples) formed a training set for training the neural network and building kriging in ArcGIS, the rest (45 samples) were the test set for testing both, the neural networks and kriging This separation was carried out randomly by using the 'create subset' in Geostatistical Analyst for ArcGIS. First, the ArcGIS application was performed to predict the values in a test data set (31 samples). In order to accomplish this goal, the ordinary kriging on the training data set (105 samples) was initially built. The predicted values in the test data set were built by the function 'Prediction' in ArcGIS.

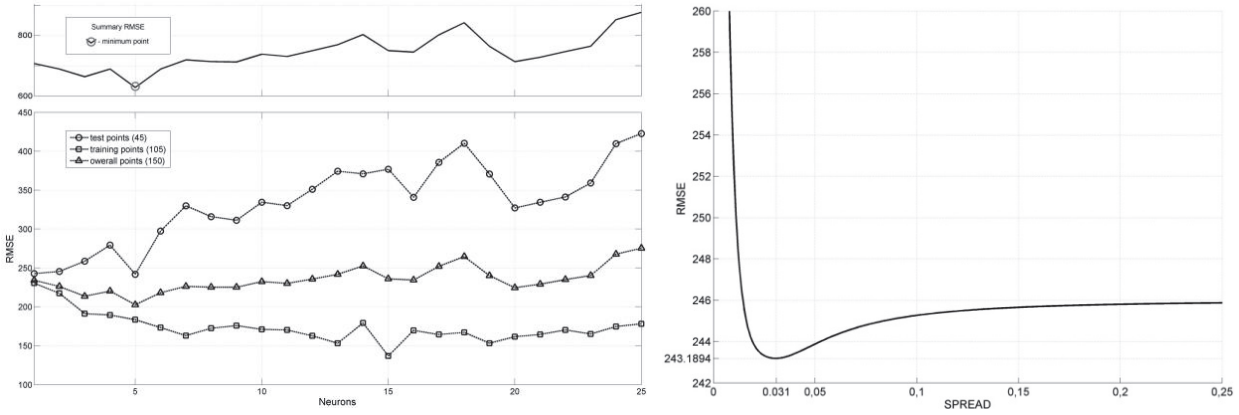
In order to assess the concentration of chromium in the training data set, two neural networks were selected: a feed-forward multilayer perceptron with the Levenberg-Marquardt training method and a generalized regression neural network. The ANNs were carried out in MATLAB using the GUI interface.

The second stage was building a MLP network. In our case, the input layer of MLP was compiled with sampling points; the hidden layer consisted of a few neurons, and the output layer representing the element content in the relevant sample. The selection of the neurons amount in the hidden layer was carried out during a specially built algorithm by the lower total RMSE of prediction of the pollutant (Cr) content for the training (105 samples), test (45 samples), and a complete set of data (150 samples). The number of neurons was varied from 2 to 25. Each selected network was trained by 500 times and the best one has been selected. Network education quality was checked by the correlation coefficient, MAE and RMSE between the result of the network prediction and data from training data set. Results of the network structure selection (number of neurons in the hidden layer) are shown in Figure 2a. The number of hidden neurons selected was 5 based on minimum of overall RMSE (see Figure 2a).

The third stage of the experiment was creating a GRNN prediction network. As it known, the configuration of GRNNs implies the distribution of the weights along the neurons of the hidden layer according the network parameter *spread*. During the simulation, the spread parameter varied from 0 to 0.3 with step 0.01, in total 300 simulations were done. The predictive accuracy of each selected network was also verified by the correlation

coefficient, MAE and RMSE between the prediction and data from training data set. Results of the GRNN structure selection are shown in Figure 2b. The minimal RMSE was achieved with spread parameter of 0.031

The final stage of the experiment was the assessment of chromium contents in points of training data set and its verification by comparing with test data set and kriging predictions. The criteria for comparison were the prediction errors MAE, RMSE and also correlation coefficients between predicted values and real ones.



(a) (b)

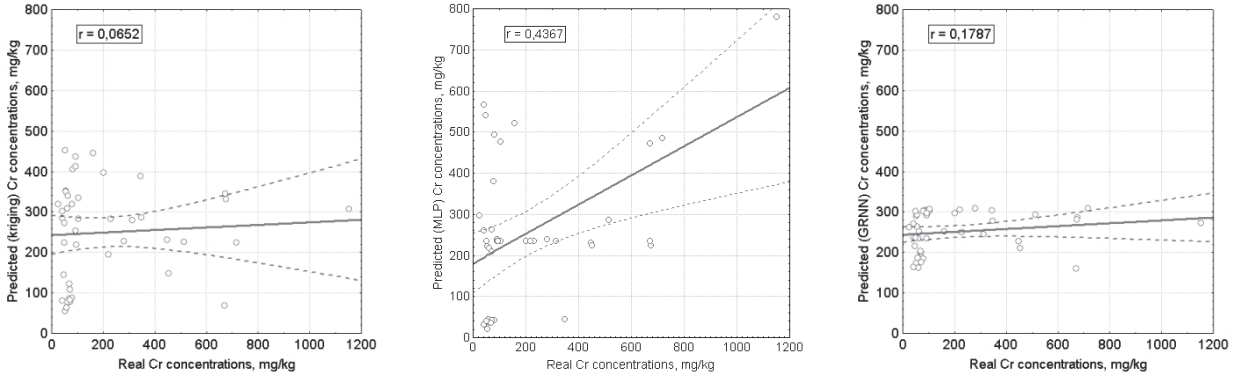
FIGURE 2. Root mean square error (RMSE) of a neural networks for test, training and overall data under different neuron number in the hidden layer for Cr: a network structure selection: a) MLP, b) GRNN

3. RESULTS AND DISCUSSION

The accuracy assessment indices of predicted concentrations are shown in Table 2 (the best values are in bold). The dependencies of predicted concentrations vs real ones are presented in Figure 3.

TABLE 2. Accuracy assessment indices of predicted concentrations of the pollutant (*Cr*)

Method	Index	Measure	Value
Kriging	MAE	mg/kg	201.88
ANN (MLP)	MAE	mg/kg	**186.15**
ANN (GRNN)	MAE	mg/kg	196.79
Kriging	RMSE	mg/kg	266.36
ANN (MLP)	RMSE	mg/kg	**241.95**
ANN (GRNN)	RMSE	mg/kg	243.20



(a) (b) (c)

FIGURE 3. Comparison of different prediction approaches, r – correlation coefficient: a) ordinary kriging, b) multilayer perceptron, c) generalized regression neural network. Dotted curves indicate confidence intervals.

Thus, we compared several approaches to modeling the distribution of the chemical elements concentrations in the surface layer of soil: geostatistical techniques (kriging) and ANN models using MLP and GRNN. The distribution of the chromium concentration in the chosen site proved irregular. A few "spots" of an abnormal distribution of chromium was found, where concentrations of the pollutant were more than 10 times higher than the levels on the rest of the polygon. Comparison of methods has shown the superiority of ANN in modeling accuracy for both networks. It was found that in contrast to expectations based on works (Mohanty & Majumdar, 1999; Shen et al., 2004, Chatterjee et al., 2007) the use of MLP gives a substantial increase in the accuracy of prediction of concentration distribution in the surface layer of soil, for abnormally distributed chromium based on RMSE (about 1% relative to GRNN and 10% relative to kriging). Moreover, ANN also provided a significantly higher correlation coefficient between real and predicted values of pollutant (Cr) content (2.5 times compared to GRNN and 6.7 times compared to kriging).

4. CONCLUSION

A study on the distribution of chromium concentrations in the surface layer of soil at the urbanized terrain of the Novy Urengoi, Yamalo-Nenets Autonomous Okrug, Russia was previously conducted. The study revealed at the investigated area an abnormal distribution of chromium formed spots with an extremely high content. The results of that study mold the basis for this work, which basic idea was to predict the element content in soil by using artificial neural network (ANN) approach. To simulate the concentrations distribution, it is proposed to use multilayer perceptron (MLP) and general regression neural network (GRNN) compared to usual geostatistical kriging modeling.

The first ANN type was a feed-forward multilayer perceptron with the Levenberg-Marquardt training algorithm with two input layer, one hidden layer and one output layer. 5 neurons in the hidden layer have been selected for modeling the distribution of chromium. The second ANN was a basic general regression neural network with selected *spread* parameter of 0.031. The frameworks of the networks were selected during specially developed simulations with hundreds steps based on minimization of RMSE.

The application of MLP gave a substantial increase in the accuracy of prediction based on RMSE (about 1% relative to GRNN and 10% relative to kriging), and also provided a significantly higher correlation coefficient between real and predicted values (2.5 times compared to GRNN and 6.7 times compared to kriging).The results showed that the ANN-based models were far more accurate than the geostatistical one (kriging). GRNN was less precise in prediction than MLP, in contrast to expectations.

The work confirms that trained ANN (in particular, MLP) is suitable for modeling an abnormal spatial distribution of pollutants. The results showed vast capabilities of ANN methods in order to improve the accuracy of modeling the spatial distribution of the contaminants concentrations in the topsoil of urban areas, which characterized by high heterogeneity. Further improvement of precision accuracy can be achieved by applying various hybrid approaches.

REFERENCES

1. Anagu I., Ingwersen J., Utermann J., Streck T. (2009) Estimation of heavy metal sorption in German soils using artificial neural networks. Geoderma. 152 . 104–112.
2. Bao, S. T., Liao, Y. M., and Hu, Y. M. (2007) Terrain interpolation based on kriging method. Geography and Geo-Information Science, 23(3): 28–32.
3. Bishop C. (1995) Neural networks for pattern recognition. Clarendon, Oxford, 504p.
4. Chatterjee S., Bandopadhyay S., Ganguli R., Bhattacherjee A., Samanta B. & Pal S. K. General regression neural network residual estimation for ore grade prediction of limestone deposit. (2007) Mining Technology, vol. 116, issue 3, 89–99.
5. Chukanov V.N., Sergeev A.P., Ovchinnikov S.M., Medvedev A.N. (2006) Diagnostics of snow-cover contamination with soluble and insoluble metal impurities, Russian Journal of Nondestructive Testing, vol. 42. 630–636.

6. Dai F., Zhoua O., Lva Z., Wang X., Liu G. (2014) Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. Ecological Indicators 45, 184–194.
7. Falamaki A. (2013) Artificial neural network application for predicting soil distribution coefficient of nickel. Journal of Environmental Radioactivity. 115. 6–12.
8. Goovaerts P. (1999) Geostatistics in soil science: State of the art and perspectives. Geoderma, 89, 1–45.
9. Guo G.H., Wu F., Xie F., Zhang R. (2012) Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China, Journal of Environmental Sciences, vol. 24, issue 3. 410–418.
10. Koike K., Matsuda S., Suzuki T. and Ohmi M. (2002). Natural Resources Research, 11, (2), 135–156.
11. Li Y., Li C., Tao J.-J., Wang L.-D. (2011) Study on Spatial Distribution of Soil Heavy Metals in Huizhou City Based on BP--ANN Modeling and GIS. Procedia Environmental Sciences 10. 1953–1960.
12. Liu Z.H., Chang Y., Chen H.W. (2008) Estimation of forest volume in Huzhong forest area based on RS, GIS and ANN (in Chinese). Chin J Appl Ecol, 19: 1891–1896.
13. Matheron G. (1963) Principles of geostatistics. Economic Geology 58: 1246–66.
14. McCulloch W.S. & Pitts W.H. (1943) A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics. Vol.5. 115–133.
15. Mohanty, K. and Majumdar, T. J. (1999) Using artificial neural networks for synthetic surface fitting and the classification of remotely sensed data. International Journal of Applied Earth Observation and Geoinformation, 1(1), 78–84.
16. Saet J.E., Revich B.A., Yanin E.P. (1990) Environment geochemistry. Nedra publishing, Moscow, Russia, 84–108 (in Russian).
17. Samanta B., Bandopadhyay S. and Ganguli R. (2004): Exploration and Mining Geology Journal, 11, 69–76.
18. Samanta B., Ganguli R. and Bandopadhyay S. (2005) Transactions of the Institution of Mining and Metallurgy, 114, 129–139.
19. Schloeder C.A., Zimmerman N.E., Jacobs M.J. (2001) Comparison of methods for interpolating soil properties using limited data. Soil Sci. Soc. Am. J. 65, 470-479.
20. Sergeev A.P., Baglaeva E.M., Shichkin A.V. (2010) Case of soil surface chromium anomaly of a northern urban territory – preliminary results, Atmospheric Pollution Research, vol. 1, 44–49.
21. Sergeev A.P., Baglaeva E.M., Medvedev A.N. The analysis of the spatial inhomogeneity of the distribution of chromium and nickel by the results of an environmental screening of a surface soil layer at residential areas of municipal formation of Novy Urengoy. GEOECOLOGY, engineering geology, hydrogeology, geocryology. 2013, Issue 3, 232-242 pp. (in Russian).
22. Sergeev A.P., Buevich A.G., Medvedev A., Subbotina I.E., Sergeeva M. (2015) Artificial neural network and kriging interpolation for the chemical elements contents in the surface layer of soil on a background area. 15th International Multidisciplinary Scientific GeoConference SGEM 2015, Conference Proceedings, 2015, Book 3 Vol. 2. 49–56.
23. Shen, Z. Q., Shi, J. B., Wang, K., et al. (2004) Neural network ensemble residual kriging application for spatial variability of soil properties. Pedosphere, 14(3), 289–296.
24. Shepherd, A.J. (1997) Second-Order Methods for Neural Networks: Fast and Reliable Training Methods for Multi-Layer Perceptrons. Springer-Verlag, 145p.
25. Sirven J.-B., Bousquet B., Canioni L., Sarger L., Tellier S., Potin-Gautier M., Le Hecho I. (2006) Qualitative and quantitative investigation of chromium-polluted soils by laser-induced breakdown spectroscopy combined with neural networks analysis. Anal Bioanal Chem. 385: 256–262.
26. Worsham L., D. Markewitz, & N. Nibbelink (2010) Incorporating spatial dependence into estimates of soil carbon contents under different land covers. Soil Sci. Am. J. 74: 635–646.
27. Yfantis E. A., Flatman, G. T. and Behar, J. V. (1987). Efficiency of kriging estimation for square, triangular, and hexagonal grids, Math. Geol., 19, 183-205.
28. Zeissler K.-O., Hertwig T. (2011) Artificial Neural Network instead of Kriging? A Case Study with Soil Contamination of Complex Sources. Landwirtschaft und Geologie, Dresden. Access 10.03.2016: http://www.beak.de/beak/sites/default/files/content/7_News/111_10_Oct_2011/Pribram2011_3.pdf
29. Zhang C., Fay D., McGrath D., Grennan E., Carton O.T. (2008) Use of trans-Gaussian kriging for national soil geochemical mapping in Ireland, Geochemistry: Exploration Environment Analysis, vol. 8, 255–265.